Clustering Longitudinal Ordinal Data via Finite Mixture of Matrix-Variate Latent Gaussians

Francesco Amato, Julien Jacques

Clustering Longitudinal Ordinal Data via Finite Mixture of Matrix-Variate Latent Gaussians

Francesco Amato [1] & Julien Jacques [1]

[1] *Univ Lyon, Univ Lyon 2, ERIC, Lyon.*
{*francesco.amato, julien.jacques*}*@univ-lyon2.fr*

Résumé. Dans les sciences sociales ou la médecine, les études sont souvent basées sur des questionnaires demandant aux participants d'exprimer des réponses ordonnées à plusieurs reprises au cours d'une période d'étude. Nous présentons un modèle pour effectuer un clustering temporel sur ces données. Le modèle repose sur un mélange de distributions normales à variation matricielle, en tenant compte simultanément des structures de dépendance temporelle internes et intermédiaires. Un algorithme MC-EM pour l'estimation du modèle est utilisé. Des applications sur les données synthétiques et réelles sont présentées.

Mots-clés. clustering. Données ordinal longitudinales. Tenseurs. Modèle de mélange. Loi gaussienne matricielle.

Abstract. In social sciences or medicine, studies are often based on questionnaires asking participants to express ordered responses several times over a study period. We present a model to perform temporal clustering on such data. The model relies on mixture of matrix-variate normal distributions, accounting for the within and between time-dependence structures simultaneously. A MC-EM algorithm for the model estimation is used. Applications on synthetic and real data are presented.

Keywords. Clustering. Ordinal longitudinal data. Three-way data. Mixture models. Matrix-variate Gaussians.

1 Context

In many areas of humanities and social sciences, the studies are based on questionnaires completed by participants several times over the study period. The researchers then analyse these questionnaires to determine typical behaviours within the studied population, being especially interested in their time evolution. Nonetheless, modelling temporal evolution is far from trivial. The most basic approach consists in performing analyses independently at each temporal phase, and then trying *a posteriori* to find links between these different analyses, by seeking from one phase to the other to find similar or different typical behaviours. An example is Selosse et al., 2019, clustering of ordinal data for an application in psychology. The ideal way to cluster these data would be to account for the

temporal evolution, modelling all the responses to the questionnaires at the same time. We propose a model-based clustering technique aiming at facilitate such temporal analysis, by grouping together the units behaving similarly in time. Over the decades, research has produced a vast number of different approaches to clustering. From our prospective, probabilistic (or model-based) clustering offers the advantage of clearly stating the assumptions behind the clustering algorithm, and allows cluster analysis to benefit from the inferential framework of statistics to address some of the practical questions arising when performing clustering: determine the number of clusters, detecting and treating outliers, assessing uncertainty in the clustering (Bouveyron et al., 2019).

1.1 Related works

An approach to clustering longitudinal data consists in arranging the data in a three-way format and modelling them through a matrix-variate mixture model. This approach offers the advantage of accounting for the overall time-behavior, grouping together the units that have a similar pattern across and within time. While not being new (Basford et al., 1985), matrix-variate distributions have recently gained attention, and mixtures of matrix-normals (MMN) have been developed and applied both in a frequentist framework in Viroli, 2011a and within a Bayesian one by Viroli, 2011b, where it was used to cluster Italian provinces based on a longitudinal crime-related score. From a frequentist point of view, these models represent a natural extension of the multivariate normal mixtures to account for temporal (or even spatial) dependencies, and have the advantage of being also relatively easy to estimate by means of EM algorithm (a nice short description of the EM application to MNN is provided in §2.1 of Wang et al., 2020). More recently, in Gallaugher et al., 2018 and Melnykov et al., 2018, 2019 extensions for non-normal skewed cases have been proposed and applied. However, matrix-variate models suffer from over-parametrization that leads to estimation issues. To overcome this issue a more parsimonious model (Sarkar et al., 2020) and a new R package (Zhu et al., 2021) has been proposed. Despite their efficacy, up to now these methods have only been applied to continous data. Our model expands the use to matrix-variate mixtures to ordinal data.

2 Model

Let denote by $y_{i,j,t}, i = 1, ..., N; j = 1, .., J$ and $t = 1, ..., T$ the observation of the j-th variable for the i-th unit at time t, that is: imagine to observe N units and measuring J different ordinal variables T times throughout the course of the study. Let us reorganize our data in a random-matrix form such that $\mathbf{Y} = \{Y_i\}_{i=1}^N$ is a sample of $J \times T$-variate matrix observations (i.e. $Y_i = (y_{i,j,t}) \in \mathbb{R}^{J \times T}$). Then, we can assume that each variable $y_{i,j,t}$ is the manifestation of an underlying latent variable $z_{i,j,t}$ which follows a Gaussian distribution, as done in the clustMD model (McParland et al., 2016). At this point, we

can assume that each observed matrix Y_i is indeed the manifestation of a latent random matrix Z_i, which follows a matrix-normal distribution.

$$Y_i = \begin{pmatrix} y_{i,1,1} & \cdots & y_{i,1,t} & \cdots & y_{i,1,T} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ y_{i,j,1} & \cdots & y_{i,j,t} & \cdots & y_{i,j,T} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ y_{i,J,1} & \cdots & y_{i,J,t} & \cdots & y_{i,J,T} \end{pmatrix} \longleftarrow Z_i = \begin{pmatrix} z_{i,1,1} & \cdots & z_{i,1,t} & \cdots & z_{i,1,T} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ z_{i,j,1} & \cdots & z_{i,j,t} & \cdots & z_{i,j,T} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ z_{i,J,1} & \cdots & z_{i,J,t} & \cdots & z_{i,J,T} \end{pmatrix}$$

To map from Y_i to Z_i, let γ_j denote a C_{j+1}-dimensional vector of thresholds that partition the real line for the j-th ordinal variable that has C_j levels and let the threshold parameters be constrained such that $-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq ... \leq \gamma_{j,C_j} = \infty$. If the latent $z_{i,j,t}$ is such that $\gamma_{j,c1} < z_{i,j,t} < \gamma_{j,c}$ then the observed ordinal response, $y_{i,j,t} = c$.

So, by assuming that each Z_i follows a matrix-normal distribution, we can then cluster our data by means of finite Mixture of Matrix-Normals (MMN) (Viroli, 2011a). Indeed, let $Z_i \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$, where $M \in \mathbb{R}^{J\times T}$ is the matrix of means, $\Phi \in \mathbb{R}^{T\times T}$ is a covariance matrix containing the variances and covariances between the T occasions or times and $\Sigma \in \mathbb{R}^{J\times J}$ is the covariance matrix containing the variance and covariances of the J variables. The matrix-normal probability density function (pdf) is given by

$$f(Z|M, \Phi, \Sigma) = (2\pi)^{-\frac{TJ}{2}} |\Phi|^{-\frac{J}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\text{tr}[\Sigma^{-1}(Z-M)\Phi^{-1}(Z-M)^\intercal]\right\}. \quad (1)$$

The matrix-normal distribution represents a natural extension of the multivariate normal distribution, since if $Z \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$, then $\text{vec}(Z) \sim \mathcal{MVN}_{JT}(\text{vec}(M), \Phi \otimes \Sigma)$, where vec(.) is the vectorization operator and $\otimes$ denotes the Kronecker product. Then, the mean and the variance of the matrix-normal distribution are:

$$\mathbb{E}(\text{vec}(Z)|M, \Phi, \Sigma) = \text{vec}(M) \quad \text{and} \quad \mathbb{V}(\text{vec}(Z)|M, \Phi, \Sigma) = \Phi \otimes \Sigma. \quad (2)$$

Being a special case of the multivariate normal distribution, the matrix-normal distribution shares the same various properties, like, for instance, closure under marginalization, conditioning and linear transformations (Gupta et al., 2000). The separability condition of the covariance matrix has the twofold advantage of allowing the modeling of the temporal pattern of interest directly on the covariance matrix Φ and of representing a more parsimonious solution than that of the unrestricted $\Phi \otimes \Sigma$.
The pdf of the MMN model is given by

$$f(Z|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \phi^{(J\times T)}(Z|M_k, \Phi_k, \Sigma_k),$$

where $\phi^{(J\times T)}$ represents the density function of a $J\times T$-dimensional matrix-variate normal, K is the number of mixture components, $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ is the vector of mixing proportions, subject to constraint $\sum_{k=1}^K \pi_k = 1$ and $\boldsymbol{\Theta} = \{\Theta_k\}_{k=1}^K$ is the set of component-specific parameters with $\Theta_k = \{M_k, \Phi_k, \Sigma_k\}$.
In addition to Z, we introduce a latent binary variable that indicate whether the unit i belongs to the k-th cluster, $\boldsymbol{\ell}_i = (\ell_{i1}, ..., \ell_{iK})$, such that $\ell_{ik} = 1$ if the i-th unit belongs to the k-th cluster. Let the binary vector $\mathbf{Y}_i^R = (Y_{i1}^R, ..., Y_{iR}^R)$ of length R indicate the observed response patter for the i-th unit, such that if the r-th pattern is observed then $Y_{ir}^R = 1$ and any other entry in the vector equals zero. We can derive the joint density of $\mathbf{Z}_i, \mathbf{Y}_i^R, \boldsymbol{\ell}_i$ as:

$$f(\mathbf{Y}_i^R, \mathbf{Z}_i, \boldsymbol{\ell}_i) = f(\mathbf{Y}_i^R|\mathbf{Z}_i, \boldsymbol{\ell}_i) f(\mathbf{Z}_i|\boldsymbol{\ell}_i) f(\boldsymbol{\ell}_i).$$

Assuming that:

$$\begin{aligned}
\boldsymbol{\ell}_i &\sim \mathcal{M}(1, \boldsymbol{\pi}),\ \boldsymbol{\pi} = (\pi_1, ..., \pi_K) \\
\mathbf{Z}_i|\boldsymbol{\ell}_{ik} = 1 &\sim \mathcal{MN}_{(J\times T)}(\mathbf{Z}_i|\Theta_k),\ \Theta_k = \{M_k, \Phi_k, \Sigma_k\}, \\
\mathbf{Y}_i^R|\mathbf{Z}_i, \boldsymbol{\ell}_{ik} = 1 &\sim \mathcal{M}(1, \boldsymbol{\xi}_i^R),\ \boldsymbol{\xi}_i^R = (\mathbf{1}_{\Omega_1}(\mathbf{Z}_i), \ldots, \mathbf{1}_{\Omega_R}(\mathbf{Z}_i))
\end{aligned}$$

we get:

$$f(\boldsymbol{\ell}_i) = \prod_{k=1}^K \pi_k^{\ell_{ik}}\ ,\ f(\mathbf{Z}_i|\boldsymbol{\ell}_i) = \prod_{k=1}^K \left[\phi^{(J\times T)}(Z_i|\Theta_k)\right]^{\ell_{ik}}\ ,\ f(\mathbf{Y}_i^R|\mathbf{Z}_i, \boldsymbol{\ell}_i) = \prod_{r=1}^R \mathbf{1}_{\Omega_r}(Z_i)^{Y_{ir}^R}\ ,$$

where $\mathcal{M}$ indicate the multinomial distribution and $\mathbf{1}_{\Omega_r}(\mathbf{Z}_i)$ is the indicator function that equals 1 when the elements in $\mathbf{Z}_i$ have values that determine the r-th pattern: one can imagine Ω_r as a $J\times T$ matrix whose elements are the indicator functions of the thresholds linked to the r-th pattern. Of course, when Z_i is given, the the value of Y_i is no more random.

3 Estimation

A key point is of course the choice of the thresholds $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^J$. In Corneli et al., 2020, thresholds are fixed in advance to avoid identifiability and computational complexity issues. While this can be a starting point in the estimation process, our model aims at treating them as parameters and estimating them, as in McParland et al., 2013, 2016.

If the thresholds are assumed to be fixed in advance, the estimation process clearly simplifies. To estimate the model, since we do not observe neither Z nor ℓ, we resort to the EM algorithm (Dempster et al., 1977). The complete log-likelihood can be then written

as

$$\log \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\Theta} | \mathbf{Y}^R, \mathbf{Z}, \boldsymbol{\ell}) \propto \sum_{i=1}^{N} \left\{ \sum_{r=1}^{R} Y_{ir} \mathbf{1}_{\Omega_r}(Z_i) + \sum_{k=1}^{K} \ell_{ik} \left[\log(\pi_k) - \frac{TJ}{2} \log(2\pi) - \frac{J}{2} \log(|\Phi_k|) - \frac{T}{2} \log(|\Sigma_k|) - \frac{1}{2} tr[\Sigma_k^{-1}(Z_i - M_k)\Phi_k^{-1}(Z_i - M_k)^\intercal] \right] \right\}$$

The expectation step (E-step) of the EM algorithm consists of computing the expectation of the complete log-likelihood with respect to the latent data Z and the cluster labels ℓ. In the first step we attempt to estimate the latent variables. For each response pattern r, we can approximate the value of $Z_i|\ell_i$ as the expected value of the truncated multivariate normal (using (2)), given the parameters Θ_k of the assigned cluster. The latent variable ℓ can be computed by means of Bayes'theorem as:

$$\mathbb{E}(\ell_{ik} | Y_i^R = r, \boldsymbol{\Theta}, \boldsymbol{\pi}) = \frac{\pi_k \int_{\Omega_r} f(Z|\Theta_k) dZ}{\sum_{k=1}^{K} \pi_k \int_{\Omega_r} f(Z|\Theta_k) dZ},$$

which requires Monte-Carlo approximation on the multivariate reparametrization. By taking the first derivatives of the log-likelihood, the M-step goes by:

$$\hat{\pi}_k = \frac{\sum_{i=1}^{N} \ell_{ik}}{N} \quad , \quad \hat{M}_k = \frac{\sum_{i=1}^{N} \ell_{ik} Z_i}{\sum_{i=1}^{N} \ell_{ik}}$$

$$\hat{\Phi}_k = \frac{\sum_{i=1}^{N} \ell_{ik} (Z_i - \hat{M}_k)^\intercal \hat{\Sigma}_k^{-1} (Z_i - \hat{M}_k)}{J \sum_{i=1}^{N} \ell_{ik}} \quad , \quad \hat{\Sigma}_k = \frac{\sum_{i=1}^{N} \ell_{ik} (Z_i - \hat{M}_k) \hat{\Phi}_k^{-1} (Z_i - \hat{M}_k)^\intercal}{T \sum_{i=1}^{N} \ell_{ik}}$$

4 Conclusions

Mixture of matrix-variate normal distributions can be an efficient way to cluster longitudinal continuous data. Assuming that ordinal variables is a discretization of a latent continuous variable allows us to extend the use of these MMN to ordinal variables. Numerical study on synthetic data sets as well as real data application concerning diet choice during the pandemic (François-Lecompte et al., 2020) will be presented.

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[2] Kaye E Basford and Geoffrey J McLachlan. "The mixture method of clustering applied to three-way data". In: *Journal of Classification* 2.1 (1985), pp. 109–125.

[3] A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 2000.

[4] Cinzia Viroli. "Finite mixtures of matrix normal distributions for classifying three-way data". In: *Statistics and Computing* 21.4 (2011), pp. 511–522.

[5] Cinzia Viroli. "Model based clustering for three-way data structures". In: *Bayesian Analysis* 6.4 (2011), pp. 573–602.

[6] Damien McParland and Isobel Claire Gormley. "Clustering ordinal data via latent variable models". In: *Algorithms from and for Nature and Life*. Springer, 2013, pp. 127–135.

[7] Damien McParland and Isobel Claire Gormley. "Model based clustering for mixed data: clustMD". In: *Advances in Data Analysis and Classification* 10.2 (2016), pp. 155–169.

[8] Michael PB Gallaugher and Paul D McNicholas. "Finite mixtures of skewed matrix variate distributions". In: *Pattern Recognition* 80 (2018), pp. 83–93.

[9] Volodymyr Melnykov and Xuwen Zhu. "On model-based clustering of skewed matrix data". In: *Journal of Multivariate Analysis* 167 (2018), pp. 181–194.

[10] Charles Bouveyron et al. *Model-based Clustering and Classification for Data Science: with applications in R*. Cambridge University Press, 2019.

[11] Volodymyr Melnykov and Xuwen Zhu. "Studying crime trends in the USA over the years 2000–2012". In: *Advances in Data Analysis and Classification* 13.1 (2019), pp. 325–341.

[12] Margot Selosse et al. "Analysing a quality of life survey using a co-clustering model for ordinal data and some dynamic implications". In: *Journal of the Royal Statistical Society: Series C Applied Statistics* 68.5 (2019), pp. 1327–1349.

[13] Marco Corneli, Charles Bouveyron, and Pierre Latouche. "Co-clustering of ordinal data via latent continuous random variables and not missing at random entries". In: *Journal of Computational and Graphical Statistics* (2020).

[14] Agnès François-Lecompte et al. "Confinement et comportements alimentaires : Quelles évolutions en matière d'alimentation durable ?" In: *Revue Française de Gestion* 293 (2020).

[15] Shuchismita Sarkar et al. "On parsimonious models for modeling matrix data". In: *Computational Statistics & Data Analysis* 142 (2020), p. 106822.

[16] Yang Wang and Volodymyr Melnykov. "On variable selection in matrix mixture modelling". In: *Stat* 9.1 (2020), e278.

[17] Xuwen Zhu, Shuchismita Sarkar, and Volodymyr Melnykov. "MatTransMix: an R Package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling". In: *Journal of Classification* (2021), pp. 1–24.